THE SEGMENTATION OF A TEXT LINE FOR A HANDWRITTEN UNCONSTRAINED DOCUMENT USING THINING ALGORITHM

Tsuruoka, S.; Adachi, Y.; Yoshikawa, T.

Published in:
EPRINTS-BOOK-TITLE

Publication date:
2004

THE SEGMENTATION OF A TEXT LINE FOR A HANDWRITTEN UNCONSTRAINED DOCUMENT USING THINNING ALGORITHM

SHINJI TSURUOKA, YUSUKE ADACHI AND TOMOHIRO YOSHIKAWA

Department of Electrical and Electronic Engineering, Faculty of Engineering,

Mie University, 1515 Kamihama, Tsu, Mie 514-8507, Japan

E-mail tsuruoka@elec.mie-u.ac.jp

For printed documents, the projection analysis of black pixels is widely used for the segmentation of a text line. However, for handwritten documents, we think that the projection analysis is not appropriate as the separating border line of a text line isn't a straight lines on a paper with no ruled line. We will extract a curved separating border line. In this paper, we propose the new segmentation of a text line from a handwritten document image using thinning algorithm. In most of documents, a text line is separated by a background region for reader to read easily. From this point of view, we use a new thinning algorithm for the background region to detect the separating border lines. In the thinned objects, the useless chains for the border line are eliminated by gradual conditions. To confirm the usefulness of this method, we applied this method to 475 text lines in 22 handwritten documents as test images and the accuracy of 90.0% is obtained.

1 Introduction

Document image processing techniques such as OCR to convert printed documents on a paper into the digital media have been used at an office. But the techniques for handwritten documents except mailing address on mail pieces, is little for Japanese documents. The most characters on a handwritten document are entered by a keyboard because OCR can't read the handwritten characters correctly. One of the reasons is that the segmentation of a text line for a document image is very difficult since the separating border line of the text line area is not a horizontal straight line [1],[2]. The typical segmentation techniques for a printed document are based on the horizontal projection analysis of the binary document image[3],[4], and the grouping of connected components [5],[6]. We think that these techniques can't use for Japanese handwritten unconstrained documents, because the separating border line of each text line isn't a straight line and the space between text lines is narrow for the separation. By these considerations, we think that the border line for a handwritten unconstrained documents should be a curved line.

In this paper, we treat the unconstrained document image on a complete white paper with no ruled lines such as the examination paper, a memorandum. We propose the new segmentation of a text line for a handwritten document image using thinning algorithm. In most of documents, a text line is separated by a background region for a reader to read easily. From this point of view, we use the

In: L.R.B. Schomaker and L.G. Vuurpijl (Eds.), Proceedings of the Seventh International Workshop on Frontiers in Handwriting Recognition, September 11-13 2000, Amsterdam, ISBN 90-76942-01-3, Nijmegen: International Unipen Foundation, pp 505-510

thinning algorithm of the background region to detect the separating border line of a text line. The thinned objects of the background are 1-pixel width line segments (chains) that are the separating border lines of the connected components (characters), and they have the useless chains for the text line segmentation. We eliminate the useless chains by gradual conditions for the property of a chain, and separate the region by the border line. The separated region becomes the input image of the word recognition system[7],[8]. To confirm the usefulness of this method, we applied this method to 472 Japanese text lines on 22 handwritten documents as a test images and the accuracy of 90.0% is obtained.

2 Text Line Segmentation Algorithm

The input document images are digitized by an image scanner at 120 dpi from unconstrained handwritten papers with no ruled line. An input gray level pixel in the document image is binarized by a threshold value into black (1: character) or white (0: background).

2.1 Limitation of thinning region

The thinning region in our method is the white region (background), and it divide one region included some text lines into some regions of each text line. The document image includes the wide white region at the margin of the top, bottom, right and left sides on the paper and some blank regions such as the separator. The limitation of thinning region is useful to reduce the calculation cost and the time. This procedure is as follows.

(1) An input binary image is divided into some small blocks by the square of $\mathbf{h_l}$ /4 x $\mathbf{h_l}$ /4 pixels, where the mean height $\mathbf{h_l}$ is determined by Fourier transform of the horizontal projection.

(2) If the block doesn't include any black pixels, then it is called the white block. If the white block is 4-connected to the image frame, then the block doesn't the thinning region. Else white blocks become the thinning regions.

2.2 Thinning process and extraction of chains

We use a new thinning algorithm based on the tsuruoka's algorithm [12]. The separating border of a text line is a closed loop, and it doesn't include end points. We delete the preserving condition for the end points in the 4-connecting thinning algorithm. The thinned image is a set of 1-pixel width line segments (border points). We classify the border points by the number of 4-neighbor border points ($\mathbf{N_4}$), that is, if the number is one, three or four ($\mathbf{N_4}$=1,3 or 4), then we define it the end point of a sequence. If both of the starting point ($\mathbf{t_1}$) and the last point ($\mathbf{t_n}$) of a sequence (border points ($\mathbf{t_1,t_2,...,t_n}$)) are end points, then the sequence is called a chain. The

chain has a set of the gradient **a** to the horizontal line and the height of the chain $\mathbf{h_c}$ (that is, vertical length) shown in Fig.1.

2.3 Gradual chain elimination

The region of a text line is the horizontal long region. The separating border line consists of the horizontal segments, and they don't have the large gradient and the large height. We assume that the useless chain has the large gradient and the large height, and the chains are gradually eliminated by the gradient **a** and the height $\mathbf{h_c}$ of the chain as the following processes.

(1) If a chain of the thinned image satisfy the first condition ($\mathbf{a} \geq 3/4$ and $\mathbf{h_c} \leq \mathbf{h_l}/2$), then it is eliminated. If the chain with the end point ($\mathbf{N_4}=1$) arises from this operation, then the chain is eliminated, too. However, the chains belonging to the image frame never eliminate for all processes.

(2) The second condition ($\mathbf{a} \geq 3/2$ and $\mathbf{h_c} \leq \mathbf{h_l}/2$) is applied for the chains of the obtained image by (1). The satisfied chains are eliminated.

(3) The third condition ($\mathbf{a} \geq 3/4$ and $\mathbf{h_c} \leq \mathbf{h_l}/4$) is applied for the chains of the obtained image by (2). The satisfied chains are eliminated.

(4) The fourth condition ($\mathbf{a} \geq 3/2$ and $\mathbf{h_c} \leq \mathbf{h_l}/4$) is applied for the chains of the obtained image by (3). The satisfied chains are eliminated.

(5) The subdivided regions are merged using the criterion in the next section.

(6) The fifth condition ($\mathbf{a} \geq 3/4$ and $\mathbf{h_c} \leq \mathbf{h_l}/4$) is applied for the chain of the obtained image by (5). The satisfied chains are eliminated.

This procedure is illustrated in Figure 2, and the elimination without this gradual procedure induces the failure of the extraction of a text line.

2.4 Merging the subdivided region

The all region is surrounded by the thinned line segments, and it involves the small subdivided region around a small connected component such as dot, comma shown in Figure 3. The subdivided region should be merged to the neighbor large region.

(1) If the height of the region $\mathbf{h_r}$ is less than the mean height of text lines $\mathbf{h_l}$ ($\mathbf{h_r} < \mathbf{h_l}$) and the width of the region $\mathbf{w_r}$ is less than the mean height of text lines $\mathbf{h_l}$ ($\mathbf{w_r} < \mathbf{h_l}$), then the regions are selected for merging .

(2) For this selected small region, the four heights $\mathbf{h_1}$ (left-upper), $\mathbf{h_2}$ (left-lower), $\mathbf{h_3}$ (right-upper) and $\mathbf{h_4}$ (right-lower) of the connected component (the black pixel) are measured, where $\mathbf{h_1}+\mathbf{h_2}=\mathbf{h_3}+\mathbf{h_4}=\mathbf{h_b}$ ($\mathbf{h_b}$:the height of the connected component). The example is shown in Figure 3.

(3) If both of upper heights ($\mathbf{h_1}$ and $\mathbf{h_3}$) or lower heights ($\mathbf{h_2}$ and $\mathbf{h_4}$) are greater than the half of the height of the connected component (($\mathbf{h_1}>\mathbf{h_b}/2$ and $\mathbf{h_3}>\mathbf{h_b}/2$) or ($\mathbf{h_2}>\mathbf{h_b}/2$ and $\mathbf{h_4}>\mathbf{h_b}/2$)), then the small region is merged to the upper or lower region. In Figure 3, the lower height $\mathbf{h_2}$ and $\mathbf{h_4}$ are greater than the half of the $\mathbf{h_b}$, and the small region is merged to the lower region.

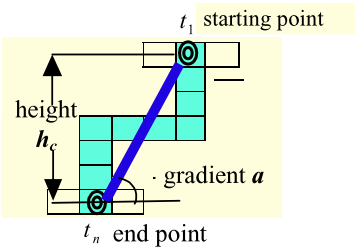


Figure 1 Property of a chain

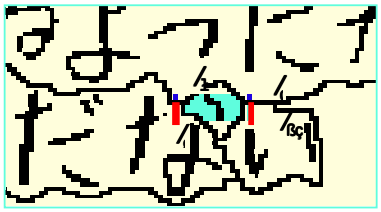
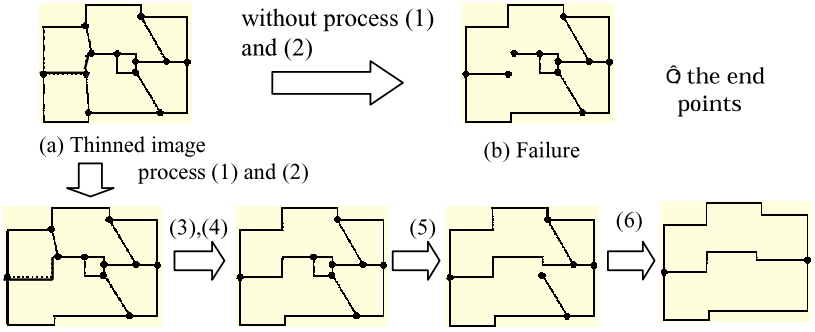


Figure 3 Merging of a subdivided region



(c) Process of the gradual elimination

Figure 2 The gradual elimination of chains

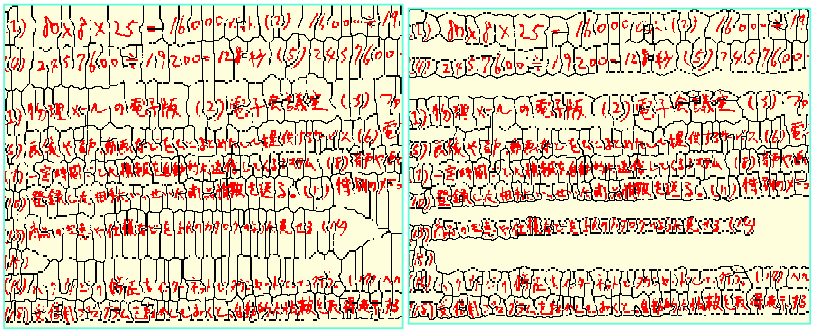
3 Experimental Results

3.1 Effect of the limitation

Examples of thinned image are shown in Figure 4. The chains in Figure 4(a) have the various gradients and heights, and the chain elimination is difficult. The comparison of the processing time about the limitation on a Pentium III (500MHz) PC with Linux is shown in Table 1. The processing time with the limitation reduced to one eight of the total time without the limitation. The separating borders extracted from Figure 4(b) is shown Figure 5. All text lines are segmented correctly.

Table 1 Times of the limitation and thinning process [s]

process	without the limitation	with the limitation
limitation	0.0	0.4
thinning	51.7	5.8
total	51.7	6.2



(a) without the limitation (b) with the limitation

Figure 4 Effect of the limitation for a thinned image

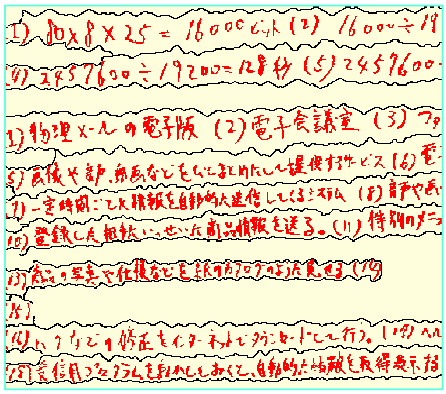


Figure 5 The resultant separating lines

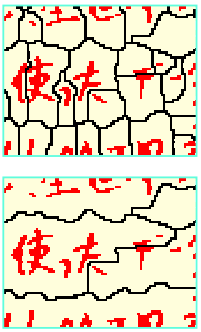


Figure 6 Subdivided text lines (upper: before the elimination, lower: after the elimination)

3.1 Results for test images

We applied this method to 472 Japanese handwritten text lines on 22 white papers as test images. The documents are the examination papers of the students, and they are written by different students. If the text line has more than one incorrect region, we treat the text line as incorrect segmentation. The number (rate) of the correct segmented text lines that is correctly separated for all components is 425 (90.0%), and the number (rate) of the subdivided text lines and non-segmented text lines are 30 (6.3%) and 18 (3.7%). The subdivided text lines shown in Figure 6 is mainly caused by the non-optimal parameters of the chain elimination threshold. We think that this error can be reduced by the optimal parameters.

4 Conclusions

We propose the new segmentation method of a text line for a handwritten unconstrained document. The features of this method are as follows.

(1) This algorithm can separate a text line even if the separating border line of the text line isn't a straight line.
(2) The limitation of thinning regions is useful for both the total processing time and the form of the thinned image.
(3) The gradual chain elimination is useful for obtaining the good separating borders.

References

1. Kimura F. , Miyake Y. and Shridhar M., Handwritten ZIP code recognition using lexicon free word recognition algorithm, Proc. of Third Int. Conf. on Document Analysis and Recognition (ICDAR'95), pp.906-910 (1995).
2. Romeo-Pakker K., Miled H. and Lecourtier Y., A new approach for latin/arabic character segmentation, Proc. of Third Int. Conf. on Document Analysis and Recognition (ICDAR'95), pp.874-877 (1995).
3. Ittner D.J. and Baird H.S., Language-free layout analysis, Proc. of second Int. Conf. on Document Analysis and Recognition (ICDAR'93), pp.336-340(1993).
4. Ha J., Haralick R. M. and Phillips I. T., Document page decomposition by the bounding-box projection technique, Proc. of Third Int. Conf. on Document Analysis and Recognition (ICDAR'95), pp.1119-1122 (1995).
5. Drivas D. and Amin A., Page segmentation and classification utilizing bottom-up approach, Proc. of Third Int. Conf. on Document Analysis and Recognition (ICDAR'95), pp.610-614 (1995).
6. Lii J. and Srihari S. N., Location of name and address on fax cover pages, Proc. of Third Int. Conf. on Document Analysis and Recognition (ICDAR'95), pp.756-759 (1995).
7. Kimura F., Shridhar M. and Chen Z., Improvements of a lexicon directed algorithm for recognition of unconstrained handwritten words, Proc. of Second Int. Conf. on Document Analysis and Recognition (ICDAR'93), pp.18-22 (1993).
8. Kimura F., Tsuruoka S., Miyake Y. and Shridhar M., A lexicon directed algorithm for recognition of unconstrained handwritten words, IEICE Trans. Inf. & Syst., Vol.E77-D, No.7, pp.785-793 (1994)
9. Tsuruoka S., Kimura F., Yoshimura M., Yokoi S. and Miyake Y., Thinning Algorithm for digital pictures and their application to handprinted characters recognition, Trans. IEICE, Vol.J66-D, No.5, pp.525-532 (1983)